PT Jurnal Cendekia Indonesia
Journal of Information Systems and Business Technology
Homepage: https://journal.jci.co.id/jisbt
Vol. 02 No. 03 (2026) Page: 785-794
E-ISSN: 3109-8886

Prediksi Customer Churn Menggunakan Decision Tree dan Random Forest dengan Pendekatan SMOTE untuk Mendukung Customer Intelligence pada Industri Telekomunikasi

[1]Nurul Fitriyah, [2]Muhammad Irfan Fauzi, [3]Mufidah Karimah

[123]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Pamulang, Tangerang Selatan, Indonesia

[1]nfitriyah878@gmail.com, [2]fauzimirfan4@gmail.com, [3]dosen02829@unpam.ac.id

Abstract

Customer churn represents a critical issue in the telecommunications sector due to its impact on customer retention levels and long-term business continuity. Consequently, companies require reliable methods to recognize customers who are likely to terminate their subscriptions. This study focuses on predicting customer churn by applying Decision Tree and Random Forest algorithms to the Telco Customer Churn dataset. The research methodology adopts the CRISP-DM framework, which encompasses data preparation, feature engineering, class balancing through the Synthetic Minority Over-sampling Technique (SMOTE), model construction, and performance evaluation. Four classification approaches were examined, including Decision Tree Gini, Decision Tree Entropy, Decision Tree Pre-Pruning, and Random Forest. Hyperparameter tuning was performed using GridSearchCV, whereas model effectiveness was assessed through Accuracy, Precision, Recall, F1-Score, and ROC-AUC metrics. The experimental results reveal that the Random Forest model produced the highest performance, achieving an accuracy of 84.88% and a ROC-AUC value of 92.94%. Furthermore, the feature importance analysis identified Contract, MonthlyCharges, and tenure as the variables with the strongest contribution to churn prediction. These findings indicate that the proposed approach can enhance Customer Intelligence by generating valuable insights into customer behavior and assisting organizations in developing more effective customer retention initiatives.

Keywords: *Customer Churn, Data Mining, Decision Tree, Random Forest, SMOTE, Customer Intelligence.*

Abstrak

Customer churn merupakan salah satu permasalahan utama yang dihadapi perusahaan telekomunikasi karena dapat menurunkan tingkat retensi pelanggan serta berdampak pada keberlanjutan bisnis. Oleh karena itu, diperlukan pendekatan yang efektif untuk mengidentifikasi pelanggan yang berpotensi menghentikan penggunaan layanan. Penelitian ini bertujuan memprediksi customer churn dengan menerapkan algoritma Decision Tree dan Random Forest pada dataset Telco Customer Churn. Tahapan penelitian mengacu pada kerangka kerja CRISP-DM yang meliputi persiapan data, rekayasa fitur, penyeimbangan data menggunakan Synthetic Minority Over-sampling Technique (SMOTE), pengembangan model, serta evaluasi kinerja model. Empat model klasifikasi yang dibandingkan dalam penelitian ini terdiri atas Decision Tree Gini, Decision Tree Entropy, Decision Tree Pre-Pruning, dan Random Forest. Proses optimasi hyperparameter dilakukan menggunakan GridSearchCV, sedangkan performa model dievaluasi berdasarkan metrik Accuracy, Precision, Recall, F1-Score, dan ROC-AUC. Hasil pengujian menunjukkan bahwa Random Forest memberikan kinerja terbaik dengan tingkat akurasi sebesar 84,88% dan nilai ROC-AUC sebesar 92,94%. Analisis feature importance mengindikasikan bahwa Contract, MonthlyCharges, dan tenure merupakan faktor yang paling berpengaruh dalam memprediksi customer churn. Temuan penelitian ini menunjukkan bahwa model yang diusulkan mampu mendukung Customer Intelligence melalui penyediaan wawasan mengenai perilaku pelanggan serta membantu perusahaan dalam merancang strategi retensi pelanggan yang lebih efektif.

Kata Kunci: Customer Churn, Data Mining, Decision Tree, Random Forest, SMOTE, Customer Intelligence.

A. PENDAHULUAN

Kemajuan teknologi informasi telah mendorong berbagai perusahaan untuk menjadikan data sebagai sumber daya strategis dalam mendukung proses pengambilan keputusan. Setiap interaksi pelanggan yang tersimpan dalam sistem menghasilkan kumpulan data yang dapat diolah menjadi informasi yang bermanfaat bagi perusahaan. Pemanfaatan data tersebut tidak terbatas pada kebutuhan operasional semata, tetapi juga berperan dalam memahami karakteristik dan perilaku pelanggan sehingga perusahaan dapat merancang strategi bisnis yang lebih tepat sasaran serta sesuai dengan kebutuhan pelanggan.

Diterima Redaksi: 01-06-2026 | Selesai Revisi: 07-06-2026 | Diterbitkan Online: 08-06-2026

Di tengah persaingan bisnis yang semakin ketat, kemampuan perusahaan dalam mempertahankan pelanggan menjadi aspek yang sangat menentukan keberhasilan usaha. Biaya yang diperlukan untuk mendapatkan pelanggan baru umumnya lebih tinggi dibandingkan biaya untuk menjaga pelanggan yang telah ada. Dengan demikian, perusahaan perlu mengidentifikasi berbagai faktor yang memengaruhi loyalitas pelanggan agar risiko kehilangan pelanggan dapat diminimalkan dan keberlangsungan bisnis dapat terjaga dalam jangka panjang.

Salah satu tantangan yang banyak ditemui perusahaan, terutama pada sektor telekomunikasi, adalah customer churn. Istilah customer churn mengacu pada kondisi ketika pelanggan menghentikan penggunaan layanan atau beralih ke penyedia layanan lainnya. Tingkat churn yang tinggi dapat menimbulkan berbagai dampak negatif, seperti penurunan pendapatan, berkurangnya pangsa pasar, serta meningkatnya pengeluaran pemasaran untuk menarik pelanggan baru. Oleh sebab itu, diperlukan suatu pendekatan yang mampu mengenali pelanggan yang berpotensi melakukan churn sehingga langkah antisipatif dapat diterapkan sejak dini.

Perkembangan teknologi analitik telah mendorong penerapan Customer Intelligence sebagai salah satu pendekatan yang banyak digunakan untuk memahami perilaku pelanggan. Pendekatan ini memanfaatkan data pelanggan sebagai sumber informasi yang dapat mendukung proses pengambilan keputusan bisnis. Dengan Customer Intelligence, perusahaan dapat memperoleh pemahaman mengenai karakteristik pelanggan, mengenali pola perilaku yang muncul, serta merancang strategi retensi yang lebih efektif berdasarkan informasi yang dihasilkan dari data.

Data Mining merupakan salah satu metode yang dapat diterapkan untuk melakukan prediksi customer churn. Melalui Data Mining, pola dan pengetahuan yang tersembunyi dalam kumpulan data berukuran besar dapat dieksplorasi dan dimanfaatkan sebagai dasar dalam pengambilan keputusan. Pada kasus customer churn, teknik klasifikasi dapat digunakan untuk membangun model yang mampu mengelompokkan pelanggan berdasarkan kemungkinan mereka untuk berhenti menggunakan layanan atau tetap menjadi pelanggan aktif. Sejumlah penelitian mengenai customer churn telah memanfaatkan berbagai algoritma klasifikasi, termasuk Decision Tree dan Random Forest. Decision Tree dikenal karena kemampuannya menghasilkan aturan klasifikasi yang sederhana sehingga mudah dipahami dan diinterpretasikan. Sementara itu, Random Forest menawarkan peningkatan kinerja prediksi melalui pendekatan ensemble learning yang mengombinasikan banyak pohon keputusan. Meskipun demikian, ketidakseimbangan kelas masih menjadi tantangan yang umum ditemukan pada dataset customer churn, karena jumlah pelanggan yang tidak melakukan churn biasanya lebih dominan dibandingkan pelanggan yang melakukan churn. Situasi tersebut dapat memengaruhi kemampuan model dalam mengenali karakteristik pelanggan churn secara optimal.

Berdasarkan kondisi tersebut, penelitian ini menerapkan Synthetic Minority Over-sampling Technique (SMOTE) sebagai metode untuk menangani ketidakseimbangan data serta melakukan perbandingan kinerja antara algoritma Decision Tree dan Random Forest dalam memprediksi customer churn. Dataset yang digunakan berupa Telco Customer Churn yang mencakup 7.043 data pelanggan telekomunikasi dengan beragam atribut yang merepresentasikan karakteristik pelanggan serta status churn.

Penelitian ini bertujuan mengembangkan model prediksi customer churn yang mampu mengidentifikasi pelanggan yang berpotensi menghentikan penggunaan layanan dengan tingkat akurasi yang lebih baik. Selain itu, penelitian ini juga diarahkan untuk menghasilkan informasi yang mendukung Customer Intelligence melalui analisis karakteristik pelanggan dan faktor-faktor yang berkontribusi terhadap terjadinya churn. Hasil penelitian diharapkan dapat menjadi dasar bagi perusahaan telekomunikasi dalam merumuskan strategi retensi pelanggan yang lebih efektif, meningkatkan tingkat loyalitas pelanggan, serta memperkuat proses pengambilan keputusan yang berbasis data.

B. TINJAUAN PUSTAKA

Customer churn mengacu pada keadaan ketika pelanggan memilih untuk menghentikan penggunaan suatu layanan atau beralih ke penyedia layanan yang berbeda. Pada sektor telekomunikasi, fenomena ini menjadi tantangan yang penting karena dapat berdampak pada penurunan pendapatan, berkurangnya loyalitas pelanggan, serta mengganggu keberlanjutan bisnis dalam jangka panjang. Oleh sebab itu, perusahaan perlu mengidentifikasi berbagai faktor yang mendorong pelanggan meninggalkan layanan agar strategi retensi dapat dirancang secara lebih efektif. Menurut (Alotaibi & Haq, 2024) penggunaan machine learning dalam prediksi customer churn dapat membantu perusahaan mengenali pelanggan yang berisiko melakukan churn sejak tahap awal, sehingga langkah pencegahan dapat diterapkan secara lebih tepat dan terarah.

Kemajuan teknologi informasi telah menghasilkan volume data pelanggan yang terus meningkat dan dapat dimanfaatkan sebagai sumber informasi untuk mendukung pengambilan keputusan bisnis. Salah satu pendekatan yang banyak digunakan dalam pengolahan data tersebut adalah Data Mining. (Han et al., 2022) menjelaskan bahwa Data Mining merupakan proses penggalian pola, hubungan, serta pengetahuan yang tersembunyi di dalam kumpulan data berukuran besar melalui berbagai metode analisis. Dalam permasalahan customer churn, teknik klasifikasi menjadi metode yang banyak diterapkan karena mampu memperkirakan apakah pelanggan akan tetap menggunakan

layanan atau memiliki kemungkinan untuk menghentikan langganannya.

Penggunaan hasil analisis data pelanggan memiliki keterkaitan yang erat dengan konsep Customer Intelligence. Pendekatan ini berfokus pada pemanfaatan informasi pelanggan untuk menghasilkan wawasan yang mendukung perumusan strategi bisnis. Melalui Customer Intelligence, perusahaan dapat memperoleh pemahaman mengenai karakteristik pelanggan, pola penggunaan layanan, preferensi yang dimiliki pelanggan, serta berbagai faktor yang memengaruhi tingkat loyalitas mereka. Wawasan tersebut selanjutnya dapat dimanfaatkan untuk menyusun program retensi yang lebih tepat sasaran dan meningkatkan kepuasan pelanggan secara berkelanjutan.

Decision Tree merupakan salah satu algoritma klasifikasi yang banyak diterapkan dalam penelitian terkait customer churn. Algoritma ini membangun model berbentuk pohon keputusan dengan memanfaatkan atribut yang memiliki pengaruh paling besar terhadap variabel target sebagai dasar pembentukan percabangan. Salah satu keunggulan utama Decision Tree terletak pada kemampuannya menghasilkan aturan klasifikasi yang mudah dipahami dan diinterpretasikan. (Ahmad et al., 2023) menyatakan bahwa Decision Tree mampu memberikan performa klasifikasi yang baik pada analisis customer churn sekaligus membantu dalam mengidentifikasi faktor-faktor yang memengaruhi keputusan pelanggan untuk tetap menggunakan layanan maupun berpindah ke layanan lain.

Selain Decision Tree, algoritma Random Forest juga banyak dimanfaatkan dalam studi prediksi customer churn. Random Forest merupakan metode ensemble learning yang mengombinasikan sejumlah pohon keputusan untuk menghasilkan prediksi yang lebih konsisten dan akurat. Pendekatan tersebut mampu meminimalkan risiko overfitting yang kerap muncul pada penggunaan satu pohon keputusan tunggal. Menurut (Breiman, 2023), Random Forest memiliki kemampuan generalisasi yang baik serta dapat menghasilkan informasi feature importance yang berguna untuk mengidentifikasi variabel-variabel yang memberikan pengaruh paling besar terhadap hasil prediksi.

Pada penerapan machine learning untuk prediksi customer churn, permasalahan ketidakseimbangan kelas (class imbalance) sering kali menjadi tantangan karena jumlah pelanggan yang tidak melakukan churn biasanya jauh lebih banyak dibandingkan pelanggan yang melakukan churn. Kondisi ini dapat menyebabkan model lebih banyak mempelajari karakteristik kelas mayoritas sehingga kemampuan dalam mendeteksi pelanggan churn menjadi kurang optimal. Untuk mengatasi permasalahan tersebut, digunakan metode Synthetic Minority Over-sampling Technique (SMOTE). (Fernandez et al., 2022) menjelaskan bahwa SMOTE bekerja dengan membentuk data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang dan kinerja model klasifikasi dapat ditingkatkan.

Berbagai penelitian terdahulu telah menunjukkan keberhasilan penerapan machine learning dalam prediksi customer churn. Penelitian yang dilakukan oleh dan (Alotaibi & Haq, 2024) menunjukkan bahwa algoritma machine learning berbasis ensemble mampu menghasilkan performa yang lebih baik dibandingkan metode klasifikasi tunggal dalam mengidentifikasi pelanggan yang berpotensi churn. Hasil penelitian tersebut membuktikan bahwa metode ensemble learning memiliki kemampuan yang lebih baik dalam menangkap pola kompleks yang terdapat pada data pelanggan.

Penelitian yang dilakukan oleh (Ahmad et al., 2023) membandingkan kinerja beberapa algoritma machine learning dalam kasus customer churn pada industri telekomunikasi. Hasil penelitian tersebut menunjukkan bahwa Random Forest mampu mencapai tingkat akurasi yang lebih baik dibandingkan sejumlah algoritma klasifikasi lainnya. Temuan ini mengindikasikan bahwa Random Forest memiliki kemampuan yang tinggi dalam menangani data pelanggan yang kompleks dan memiliki karakteristik yang beragam.

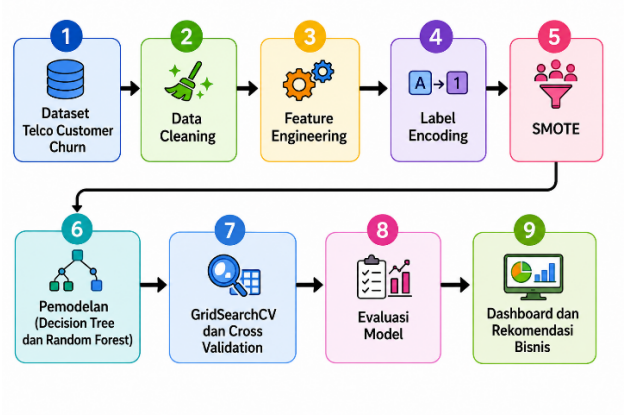
Pada penelitian berikutnya, (Ullah, 2022) menerapkan metode SMOTE untuk mengatasi masalah ketidakseimbangan kelas yang terdapat pada dataset customer churn. Hasil yang diperoleh menunjukkan bahwa penggunaan SMOTE mampu meningkatkan nilai recall dan F1-score, sehingga model menjadi lebih efektif dalam mengenali pelanggan yang berpotensi melakukan churn. Temuan tersebut menegaskan bahwa proses penyeimbangan data memegang peranan penting dalam meningkatkan performa model klasifikasi.

Penelitian yang dilakukan oleh (Huang et al., 2023) juga mengungkapkan bahwa beberapa faktor, seperti jenis kontrak pelanggan, biaya layanan bulanan, dan durasi berlangganan, memberikan pengaruh yang signifikan terhadap terjadinya customer churn. Dengan memanfaatkan analisis feature importance pada algoritma Random Forest, penelitian tersebut berhasil mengidentifikasi variabel-variabel utama yang berkontribusi terhadap keputusan pelanggan untuk menghentikan penggunaan layanan. Hasil tersebut semakin memperkuat pentingnya penerapan Customer Intelligence dalam mendukung proses pengambilan keputusan yang berbasis data.

Berdasarkan berbagai teori dan hasil penelitian terdahulu, dapat disimpulkan bahwa penerapan Data Mining yang dipadukan dengan algoritma Decision Tree, Random Forest, serta metode SMOTE merupakan pendekatan yang efektif untuk memprediksi customer churn. Selain mampu menghasilkan model prediksi dengan tingkat akurasi yang baik, kombinasi tersebut juga dapat memberikan informasi mengenai faktor-faktor yang memengaruhi perilaku pelanggan. Informasi tersebut dapat dimanfaatkan untuk mendukung implementasi Customer Intelligence sekaligus membantu perusahaan dalam merancang strategi retensi pelanggan yang lebih efektif.

C. METODE

Penelitian ini menerapkan pendekatan data mining untuk melakukan prediksi customer churn pada pelanggan telekomunikasi dengan memanfaatkan dataset Telco Customer Churn. Dataset tersebut mencakup 7.043 data pelanggan yang terdiri atas berbagai atribut yang merepresentasikan karakteristik pelanggan serta informasi mengenai status churn.



Gambar 1. Alur Penelitian

Berdasarkan Gambar 1, tahapan penelitian dimulai dengan proses pengumpulan dataset Telco Customer Churn. Selanjutnya, data diproses melalui tahap data cleaning untuk menangani missing value serta memastikan kesesuaian tipe data pada setiap atribut. Setelah proses tersebut, dilakukan feature engineering dan transformasi data menggunakan metode Label Encoding. Untuk menangani permasalahan ketidakseimbangan kelas, diterapkan teknik SMOTE. Data yang telah melalui seluruh tahapan prapemrosesan kemudian digunakan dalam proses pembangunan model menggunakan algoritma Decision Tree dan Random Forest. Kinerja model yang dihasilkan selanjutnya dievaluasi dengan beberapa metrik evaluasi, kemudian hasil analisis disajikan dalam bentuk dashboard visualisasi serta rekomendasi bisnis.

Tabel 1. Tahapan Penelitian

Tahap	Aktivitas
Data Collection	Menggunakan dataset Telco Customer Churn
Data Cleaning	Menghapus customerID dan menangani missing value
Feature Engineering	Membuat AvgChargePerMonth dan IsNewCustomer
Data Transformation	Label Encoding
Data Balancing	SMOTE
Modeling	Decision Tree dan Random Forest
Hyperparameter Tuning	GridSearchCV
Evaluation	Accuracy, Precision, Recall, F1-Score, ROC-AUC
Visualization	Dashboard Customer Churn

Berdasarkan Tabel 1, penelitian ini dilaksanakan melalui sembilan tahapan utama yang dimulai dari pengumpulan data dan berakhir pada penyajian hasil dalam bentuk visualisasi. Pada tahap data preparation dilakukan serangkaian proses yang meliputi data cleaning, feature engineering, transformasi data, serta penyeimbangan distribusi kelas menggunakan metode SMOTE. Tahap berikutnya adalah pembangunan model klasifikasi dengan algoritma Decision Tree dan Random Forest yang dioptimalkan menggunakan GridSearchCV. Selanjutnya, penelitian diakhiri dengan proses evaluasi model dan penyajian hasil analisis dalam bentuk dashboard sebagai pendukung pengambilan keputusan bisnis.

Proses evaluasi model bertujuan untuk menilai kemampuan algoritma dalam melakukan klasifikasi customer churn. Dalam penelitian ini, pengukuran performa model dilakukan menggunakan beberapa metrik evaluasi, yaitu Accuracy, Precision, Recall, dan F1-Score. Penggunaan metrik tersebut dipilih karena dapat memberikan gambaran yang komprehensif mengenai tingkat kinerja dan ketepatan model dalam mengidentifikasi pelanggan yang berpotensi melakukan churn.

1. Accuracy

Accuracy merupakan metrik yang digunakan untuk menilai tingkat keberhasilan model dalam mengklasifikasikan seluruh data secara tepat. Nilai accuracy diperoleh dengan menghitung proporsi prediksi yang benar terhadap keseluruhan data yang diuji, sebagaimana ditunjukkan pada persamaan berikut.

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

Keterangan:

- TP (*True Positive*) = jumlah data churn yang berhasil diprediksi churn.
- TN (*True Negative*) = jumlah data tidak churn yang berhasil diprediksi tidak churn.
- FP (*False Positive*) = jumlah data tidak churn yang diprediksi churn.
- FN (*False Negative*) = jumlah data churn yang diprediksi tidak churn.

2. Precision

Precision merupakan metrik yang digunakan untuk mengevaluasi tingkat ketepatan model dalam mengidentifikasi pelanggan yang diprediksi termasuk ke dalam kategori churn. Nilai precision diperoleh berdasarkan perbandingan antara jumlah prediksi churn yang benar dengan seluruh data yang diprediksi sebagai churn, sebagaimana ditunjukkan pada persamaan berikut.

$$Precision = \frac{TP}{TP + FP}$$

3. Recall

Recall merupakan metrik yang digunakan untuk mengukur kemampuan model dalam mengenali pelanggan yang benar-benar termasuk dalam kategori churn. Nilai recall diperoleh dari perbandingan antara jumlah pelanggan churn yang berhasil teridentifikasi dengan seluruh pelanggan yang sebenarnya mengalami churn, sebagaimana ditunjukkan pada persamaan berikut.

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score

F1-Score merupakan metrik evaluasi yang digunakan untuk menilai keseimbangan antara nilai precision dan recall dalam suatu model klasifikasi. Metrik ini memberikan gambaran yang lebih menyeluruh mengenai kinerja model, terutama ketika terdapat ketidakseimbangan kelas pada data. Nilai F1-Score dihitung berdasarkan persamaan berikut.

$$F1 - Score\ 2\ \times = \frac{Precision \times Recall}{Precision + Recall}$$

5. ROC-AUC

ROC-AUC (Receiver Operating Characteristic - Area Under Curve) merupakan metrik evaluasi yang digunakan untuk menilai kemampuan model dalam membedakan pelanggan yang melakukan churn dan pelanggan yang tidak melakukan churn. Nilai ROC-AUC yang semakin mendekati 1 menunjukkan bahwa model memiliki kemampuan klasifikasi yang semakin baik. Dalam penelitian ini, ROC-AUC digunakan sebagai salah satu indikator evaluasi untuk membandingkan kinerja algoritma Decision Tree dan Random Forest.

Berdasarkan rangkaian tahapan penelitian yang telah dilakukan, dihasilkan model prediksi customer churn menggunakan algoritma Decision Tree dan Random Forest. Hasil dari proses pemodelan beserta evaluasinya selanjutnya dibahas pada bagian hasil dan pembahasan guna menganalisis performa masing-masing model dalam mengidentifikasi pelanggan yang berpotensi melakukan churn.

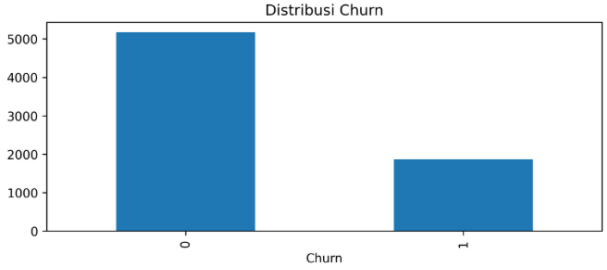
D. HASIL DAN PEMBAHASAN

Bagian ini menguraikan hasil penerapan algoritma Decision Tree dan Random Forest untuk memprediksi customer churn pada dataset Telco Customer Churn. Penelitian dilaksanakan melalui sejumlah tahapan, yaitu persiapan data, penyeimbangan kelas menggunakan metode SMOTE, pembangunan model klasifikasi, optimasi hyperparameter dengan GridSearchCV, serta evaluasi kinerja model menggunakan berbagai metrik pengukuran. Selain proses pemodelan, penelitian ini juga mencakup analisis feature importance untuk mengidentifikasi faktor-faktor yang berpengaruh terhadap customer churn. Dengan demikian, hasil yang diperoleh tidak hanya berupa model prediksi, tetapi juga informasi yang dapat dimanfaatkan dalam mendukung penerapan Customer Intelligence. Pembahasan pada bagian ini berfokus pada interpretasi hasil analisis, perbandingan performa model, serta implikasi temuan penelitian yang dapat digunakan sebagai dasar pengambilan keputusan bisnis.

1. Hasil Persiapan Data

Penelitian ini memanfaatkan dataset Telco Customer Churn yang berisi 7.043 data pelanggan dengan sejumlah atribut yang menggambarkan karakteristik pelanggan serta status churn. Sebelum dilakukan proses pemodelan, data terlebih dahulu melalui tahap data preparation yang mencakup pembersihan data, penanganan missing value, rekayasa fitur, transformasi data, dan penyeimbangan kelas. Berdasarkan hasil pemeriksaan awal, ditemukan 11 nilai yang hilang pada atribut TotalCharges. Permasalahan tersebut diatasi dengan menerapkan metode mean imputation, yaitu menggantikan nilai yang kosong menggunakan nilai rata-rata dari atribut yang bersangkutan. Dengan pendekatan tersebut, seluruh data tetap dapat digunakan dalam proses analisis tanpa mengurangi jumlah observasi yang tersedia. Setelah proses pembersihan data selesai dilakukan, tidak ditemukan lagi missing value pada dataset sehingga data dinyatakan siap untuk digunakan pada tahap pemodelan.

Tahap berikutnya adalah analisis distribusi kelas untuk mengetahui proporsi pelanggan yang melakukan churn dan yang tidak melakukan churn. Hasil analisis menunjukkan bahwa jumlah pelanggan yang tidak melakukan churn sebanyak 5.174 pelanggan, sedangkan pelanggan yang melakukan churn berjumlah 1.869 pelanggan. Perbedaan proporsi tersebut menunjukkan adanya kondisi ketidakseimbangan kelas (class imbalance) yang berpotensi membuat model lebih banyak mempelajari pola dari kelas mayoritas. Untuk mengatasi kondisi tersebut, diterapkan metode Synthetic Minority Over-sampling Technique (SMOTE) yang menghasilkan data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang. Setelah proses SMOTE diterapkan, jumlah data bertambah menjadi 10.348 data yang selanjutnya digunakan dalam proses pelatihan dan pengujian model klasifikasi.

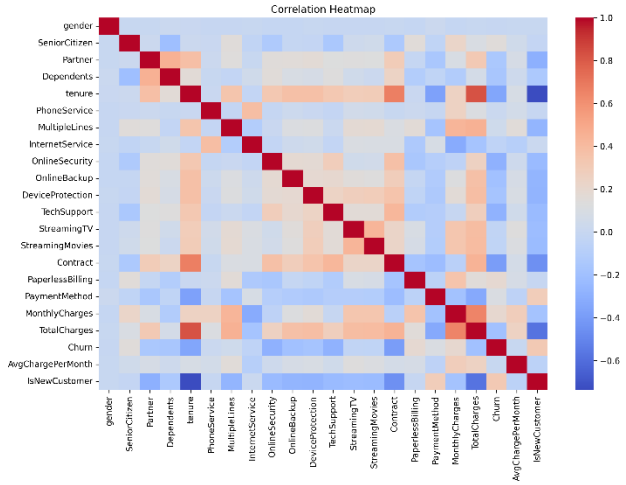


Gambar 2. Distribusi *Customer Churn*

Berdasarkan Gambar 2, hasil analisis distribusi data menunjukkan bahwa jumlah pelanggan yang tidak melakukan churn mencapai 5.174 pelanggan, sedangkan pelanggan yang melakukan churn hanya berjumlah 1.869 pelanggan. Perbedaan jumlah yang cukup signifikan antara kedua kelas tersebut menunjukkan adanya kondisi class imbalance yang berpotensi memengaruhi performa model, karena model cenderung lebih banyak mempelajari karakteristik dari kelas mayoritas dibandingkan kelas minoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan metode Synthetic Minority Over-sampling Technique (SMOTE) yang menghasilkan data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih seimbang. Setelah proses SMOTE dilakukan, jumlah data meningkat menjadi 10.348 data yang kemudian dimanfaatkan dalam proses pelatihan dan pengujian model klasifikasi.

2. Analisis Korelasi Variabel

Sebelum memasuki tahap pembangunan model, dilakukan analisis korelasi untuk mengidentifikasi pola hubungan yang terdapat di antara variabel-variabel dalam dataset. Analisis ini bertujuan memberikan pemahaman awal mengenai tingkat keterkaitan masing-masing atribut terhadap variabel target (Churn), sekaligus mengevaluasi hubungan yang terjadi antarvariabel independen. Hasil yang diperoleh dari analisis korelasi dapat digunakan untuk memahami karakteristik data secara lebih mendalam, mengidentifikasi variabel yang berpotensi memengaruhi terjadinya churn, serta mendukung proses pengembangan model klasifikasi agar dapat menghasilkan kinerja yang lebih optimal.



Gambar 3. *Correlation Heatmap*

Berdasarkan Gambar 3, mayoritas variabel memiliki tingkat korelasi pada kategori rendah hingga sedang, yang menunjukkan bahwa setiap atribut berkontribusi melalui informasi yang berbeda dan saling melengkapi dalam menggambarkan karakteristik pelanggan. Hubungan positif yang relatif kuat terlihat pada pasangan variabel tenure dan TotalCharges serta MonthlyCharges dan TotalCharges. Kondisi tersebut mengindikasikan bahwa peningkatan masa berlangganan maupun biaya bulanan cenderung diikuti oleh peningkatan total pembayaran pelanggan.

Di sisi lain, variabel Contract, tenure, OnlineSecurity, dan TechSupport menunjukkan hubungan negatif terhadap Churn. Hal tersebut mengisyaratkan bahwa pelanggan yang memiliki kontrak dengan durasi lebih panjang, telah berlangganan dalam waktu yang lebih lama, serta memanfaatkan layanan keamanan dan dukungan teknis cenderung memiliki kemungkinan yang lebih rendah untuk melakukan churn. Sebaliknya, variabel MonthlyCharges memperlihatkan hubungan positif dengan Churn, yang menunjukkan bahwa semakin tinggi biaya layanan yang dibebankan kepada pelanggan, semakin besar pula potensi pelanggan untuk menghentikan penggunaan layanan. Temuan ini memberikan gambaran awal mengenai faktor-faktor yang berkaitan dengan customer churn dan menjadi landasan dalam proses pengembangan model klasifikasi.

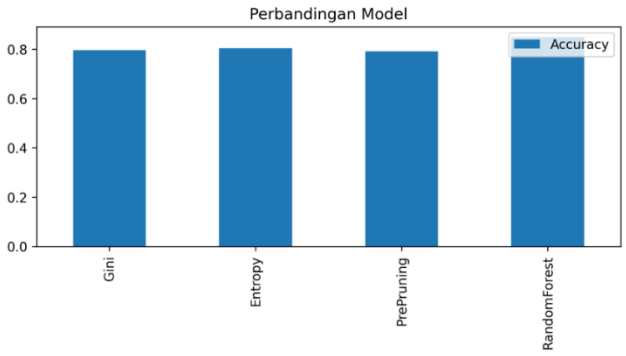
3. Hasil Pemodelan dan Perbandingan Model

Tahap pemodelan dilakukan dengan tujuan membangun model klasifikasi yang dapat memperkirakan kemungkinan pelanggan melakukan churn. Dalam penelitian ini, empat model klasifikasi diterapkan, yaitu Decision Tree dengan kriteria Gini Index, Decision Tree dengan kriteria Entropy, Decision Tree Pre-Pruning, serta Random Forest. Setiap model kemudian dievaluasi berdasarkan nilai accuracy untuk mengukur tingkat ketepatan prediksi dan menentukan model yang memiliki performa terbaik dalam mengidentifikasi pelanggan yang berpotensi melakukan churn.

Tabel 2. Perbandingan Performa Model

Model	Accuracy
Decision Tree Gini	79,61%
Decision Tree Entropy	80,39%
Decision Tree Pre-Pruning	79,08%
Random Forest	84,88%

Berdasarkan Tabel 2, algoritma Random Forest menghasilkan tingkat akurasi tertinggi, yaitu sebesar 84,88%. Di bawahnya, Decision Tree Entropy memperoleh akurasi sebesar 80,39%, diikuti oleh Decision Tree Gini dengan nilai 79,61%, serta Decision Tree Pre-Pruning yang mencapai 79,08%. Hasil tersebut menunjukkan bahwa seluruh model memiliki kemampuan yang cukup baik dalam melakukan klasifikasi customer churn. Namun demikian, Random Forest menunjukkan kinerja yang paling unggul dibandingkan ketiga model lainnya berdasarkan nilai akurasi yang diperoleh.



Gambar 4. Perbandingan Akurasi Model

Gambar 4 menunjukkan perbandingan tingkat akurasi masing-masing model secara visual. Berdasarkan grafik tersebut, Random Forest memperoleh nilai akurasi tertinggi dibandingkan seluruh model Decision Tree yang diuji. Kinerja yang lebih baik ini dipengaruhi oleh penggunaan pendekatan ensemble learning yang mengombinasikan banyak pohon keputusan untuk menghasilkan prediksi akhir. Melalui pendekatan tersebut, risiko terjadinya overfitting dapat diminimalkan, kemampuan generalisasi model menjadi lebih baik, dan hasil prediksi yang dihasilkan cenderung lebih konsisten terhadap data yang belum pernah digunakan sebelumnya. Berdasarkan hasil pengujian dan perbandingan performa yang telah dilakukan, Random Forest ditetapkan sebagai model terbaik dan digunakan pada tahap evaluasi lanjutan karena menunjukkan kemampuan klasifikasi yang paling optimal pada dataset Telco Customer Churn.

4. Optimasi dan Evaluasi Model Random Forest

Setelah model terbaik berhasil ditentukan, tahap selanjutnya adalah melakukan optimasi hyperparameter menggunakan metode GridSearchCV. Proses ini bertujuan untuk menemukan kombinasi parameter yang paling sesuai sehingga model dapat mencapai performa yang lebih optimal dalam melakukan prediksi customer churn.

Tabel 3. Paramater Terbaik *Random Forest*

Parameter	Nilai
n_estimators	200
max_depth	15
min_samples_split	2

Berdasarkan hasil proses optimasi, kombinasi hyperparameter terbaik diperoleh pada nilai n_estimators = 200, max_depth = 15, dan min_samples_split = 2. Kombinasi parameter tersebut selanjutnya diterapkan pada tahap pelatihan model akhir untuk menghasilkan performa prediksi yang optimal.

Untuk mengevaluasi tingkat kestabilan model, dilakukan pengujian menggunakan metode 5-Fold Cross Validation.

Tabel 4. Hasil *Cross Validation*

Fold	Accuracy
Fold 1	74,98%
Fold 2	78,21%
Fold 3	88,26%
Fold 4	89,32%
Fold 5	89,99%
Rata-rata	84,15%

Nilai rata-rata cross validation sebesar 84,15% menunjukkan bahwa model memiliki kemampuan generalisasi yang baik dalam menangani data yang belum pernah digunakan sebelumnya. Hasil tersebut mengindikasikan bahwa model tidak hanya mampu memberikan kinerja yang baik pada data pelatihan, tetapi juga dapat mempertahankan tingkat performanya ketika diterapkan pada data baru.

5. Evaluasi Klasifikasi Random Forest

Selain menggunakan metrik accuracy, evaluasi performa model juga dilakukan dengan memanfaatkan precision, recall, F1-score, ROC-AUC, serta classification report. Penggunaan berbagai metrik tersebut bertujuan untuk memberikan penilaian yang lebih menyeluruh terhadap kemampuan model dalam memprediksi customer churn, sehingga kinerja model dapat dianalisis secara lebih komprehensif dari berbagai aspek klasifikasi.

Tabel 5. Hasil Evaluasi Random Forest

Metrik	Nilai
Accuracy	84,88%
Precision	83,24%
Recall	87,34%
F1-Score	85,24%
ROC-AUC	92,94%

Berdasarkan Tabel 5, model Random Forest memperoleh nilai accuracy sebesar 84,88%, yang mengindikasikan bahwa sebagian besar data pelanggan dapat diklasifikasikan dengan tepat. Nilai precision sebesar 83,24% menunjukkan bahwa sebagian besar pelanggan yang diprediksi akan melakukan churn memang termasuk dalam kategori churn yang sebenarnya. Di sisi lain, nilai recall sebesar 87,34% mengindikasikan bahwa model memiliki kemampuan yang baik dalam mendeteksi pelanggan yang berpotensi menghentikan penggunaan layanan.

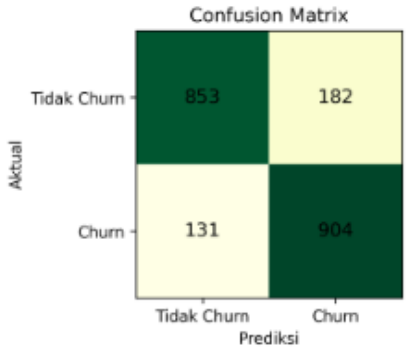
Nilai F1-Score yang mencapai 85,24% menunjukkan adanya keseimbangan yang baik antara precision dan recall dalam proses klasifikasi. Selain itu, nilai ROC-AUC sebesar 92,94% mengindikasikan bahwa model memiliki kemampuan yang sangat baik dalam membedakan pelanggan yang melakukan churn dan pelanggan yang tidak melakukan churn.

Tabel 6. *Classification Report Random Forest*

Kelas	Precision	Recall	F1-Score
Tidak Churn	0,87	0,82	0,84
Churn	0,83	0,87	0,85

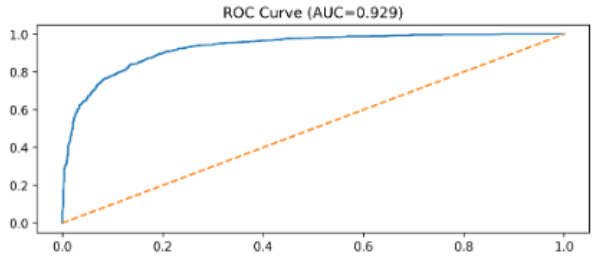
Berdasarkan Tabel 6, model memperlihatkan kinerja yang relatif seimbang dalam mengklasifikasikan kedua kelas. Nilai recall sebesar 87% pada kelas churn menunjukkan bahwa model mampu mengidentifikasi sebagian besar

pelanggan yang berpotensi menghentikan penggunaan layanan. Hasil tersebut memiliki peran penting karena tujuan utama penelitian ini adalah mendeteksi pelanggan yang berisiko melakukan churn, sehingga perusahaan dapat menerapkan strategi retensi dan langkah pencegahan secara lebih dini.



Gambar 5. *Confusion Matrix Random Forest*

Berdasarkan Gambar 5, model berhasil mengklasifikasikan dengan benar sebanyak 853 pelanggan pada kategori tidak churn dan 904 pelanggan pada kategori churn. Meskipun masih terdapat beberapa kesalahan prediksi yang termasuk dalam kategori false positive dan false negative, jumlahnya relatif lebih sedikit dibandingkan dengan jumlah prediksi yang benar. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang baik dalam membedakan pelanggan churn dan tidak churn, sehingga mampu menghasilkan kinerja klasifikasi yang efektif pada kedua kelas.

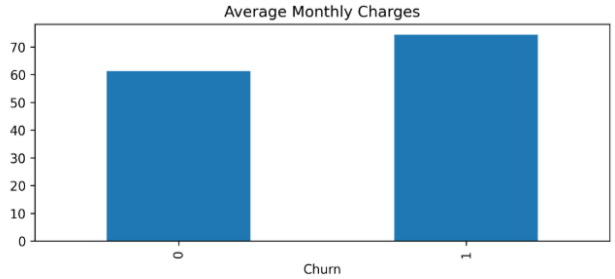


Gambar 6. *ROC Curve Random Forest*

Kurva ROC yang ditampilkan pada Gambar 6 memperlihatkan bahwa model memiliki kemampuan diskriminasi yang sangat baik, yang ditunjukkan oleh nilai AUC sebesar 0,9294. Posisi kurva yang berada jauh di atas garis diagonal mengindikasikan bahwa model mampu membedakan pelanggan yang melakukan churn dan pelanggan yang tidak melakukan churn secara efektif, sehingga menghasilkan performa klasifikasi yang tinggi.

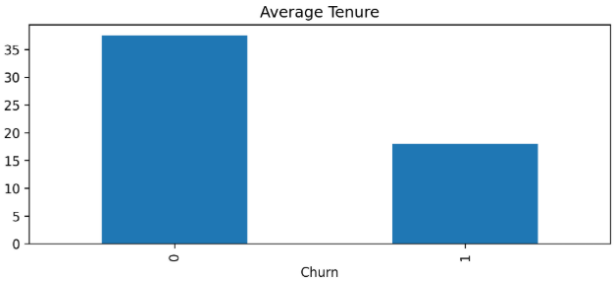
6. Analisis Karakteristik Pelanggan untuk Customer Intelligence

Selain mengembangkan model prediksi customer churn, penelitian ini juga melakukan analisis terhadap karakteristik pelanggan yang berhubungan dengan terjadinya churn guna mendukung penerapan Customer Intelligence. Analisis tersebut bertujuan menghasilkan wawasan yang dapat membantu perusahaan memahami perilaku pelanggan serta mengidentifikasi faktor-faktor yang berpengaruh terhadap keputusan pelanggan untuk tetap menggunakan layanan atau menghentikannya.



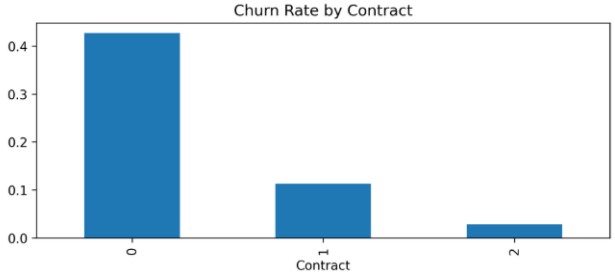
Gambar 7. *Average Monthly Charges Berdasarkan Status Churn*

Berdasarkan Gambar 7, rata-rata biaya bulanan yang dibayarkan oleh pelanggan yang melakukan churn terlihat lebih tinggi dibandingkan pelanggan yang tetap menggunakan layanan. Temuan tersebut mengindikasikan bahwa besarnya biaya layanan dapat menjadi salah satu faktor yang berkontribusi terhadap keputusan pelanggan untuk menghentikan penggunaan layanan.



Gambar 8. *Average Tenure* Berdasarkan Status *Churn*

Berdasarkan Gambar 8, pelanggan yang tidak melakukan churn memiliki rata-rata durasi berlangganan yang lebih lama dibandingkan pelanggan yang melakukan churn. Temuan tersebut mengindikasikan bahwa tingkat loyalitas pelanggan cenderung meningkat seiring dengan bertambahnya masa penggunaan layanan, sehingga pelanggan yang telah berlangganan lebih lama memiliki kecenderungan yang lebih rendah untuk menghentikan penggunaan layanan.



Gambar 9. *Churn Rate* Berdasarkan Jenis Kontrak

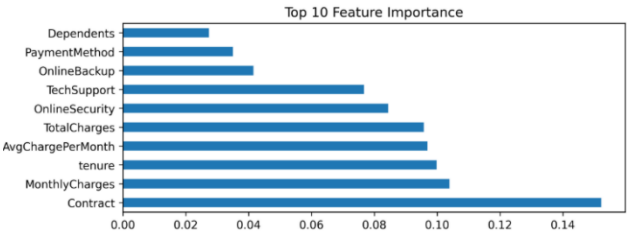
Berdasarkan Gambar 9, tingkat churn tertinggi ditemukan pada pelanggan yang menggunakan kontrak bulanan jika dibandingkan dengan pelanggan yang memiliki kontrak satu tahun maupun dua tahun. Hasil tersebut mengindikasikan bahwa jenis kontrak memiliki peranan yang signifikan dalam memengaruhi keputusan pelanggan untuk mempertahankan atau menghentikan penggunaan layanan.

7. Analisis Feature Importance

Tabel 7. Top 10 Feature Importance

Fitur	Importance
Contract	0,1522
MonthlyCharges	0,1039
tenure	0,0999
AvgChargePerMonth	0,0969
TotalCharges	0,0958
OnlineSecurity	0,0844
TechSupport	0,0767
OnlineBackup	0,0415
PaymentMethod	0,0350
Dependents	0,0274

Berdasarkan hasil analisis feature importance, variabel Contract menjadi faktor yang memberikan pengaruh terbesar terhadap customer churn dengan nilai importance sebesar 0,1522. Variabel tersebut kemudian diikuti oleh MonthlyCharges, tenure, AvgChargePerMonth, dan TotalCharges sebagai faktor-faktor dengan kontribusi tertinggi terhadap hasil prediksi. Temuan ini menunjukkan bahwa jenis kontrak, besarnya biaya layanan, serta durasi berlangganan merupakan faktor utama yang berperan dalam memengaruhi tingkat loyalitas pelanggan.



Gambar 10. Top 10 Feature Importance

Berdasarkan Gambar 10, variabel Contract memperoleh nilai feature importance tertinggi dibandingkan seluruh variabel lainnya. Temuan tersebut mengindikasikan bahwa jenis kontrak menjadi faktor yang paling dominan dalam memprediksi customer churn. Selain itu, variabel MonthlyCharges, tenure, AvgChargePerMonth, dan TotalCharges juga menunjukkan kontribusi yang signifikan terhadap performa model. Di sisi lain, variabel OnlineSecurity dan TechSupport turut memberikan pengaruh dalam membedakan pelanggan yang melakukan churn dan yang tidak melakukan churn. Hasil analisis ini menunjukkan bahwa jenis kontrak, biaya layanan, dan durasi berlangganan merupakan faktor-faktor utama yang berperan dalam memengaruhi perilaku pelanggan.

8. Implikasi Customer Intelligence

Hasil analisis karakteristik pelanggan serta feature importance menunjukkan bahwa terdapat sejumlah faktor yang memberikan pengaruh signifikan terhadap terjadinya customer churn. Informasi yang diperoleh dari analisis tersebut dapat digunakan oleh perusahaan sebagai landasan dalam merancang strategi retensi pelanggan yang lebih efektif, terarah, dan sesuai dengan karakteristik pelanggan yang berpotensi melakukan churn.

Tabel 8. Rekomendasi Strategi Retensi Pelanggan

Hasil Analisis	Strategi Retensi
Contract merupakan faktor paling berpengaruh terhadap churn	Memfokuskan program retensi pada pelanggan kontrak bulanan
MonthlyCharges memiliki pengaruh tinggi terhadap churn	Memberikan promo atau penawaran khusus bagi pelanggan dengan biaya layanan tinggi
Pelanggan dengan tenure rendah lebih rentan churn	Menerapkan program loyalitas untuk meningkatkan retensi pelanggan baru

Berdasarkan Tabel 8, pelanggan yang menggunakan kontrak bulanan perlu menjadi fokus utama dalam strategi retensi karena kelompok ini memiliki tingkat kecenderungan churn yang lebih tinggi dibandingkan pelanggan dengan kontrak jangka panjang. Selain itu, pelanggan yang menanggung biaya layanan bulanan yang relatif tinggi juga memerlukan perhatian khusus melalui pemberian promosi atau penawaran yang lebih relevan dengan kebutuhan dan preferensi mereka.

Di sisi lain, pelanggan dengan masa berlangganan yang masih relatif singkat menunjukkan tingkat risiko churn yang lebih besar dibandingkan pelanggan yang telah menggunakan layanan dalam jangka waktu yang lebih lama. Oleh sebab itu, perusahaan dapat menerapkan program loyalitas serta meningkatkan kualitas layanan sejak tahap awal penggunaan untuk mendorong kepuasan dan mempertahankan pelanggan.

Secara umum, hasil penelitian ini menunjukkan bahwa model prediksi customer churn dapat dimanfaatkan sebagai pendukung implementasi Customer Intelligence dalam mengidentifikasi pelanggan yang berpotensi melakukan churn. Dengan demikian, perusahaan dapat mengambil langkah retensi secara lebih cepat, tepat sasaran, dan berbasis data.

Simpulan

Berdasarkan hasil penelitian yang telah dilakukan, metode Decision Tree dan Random Forest berhasil diterapkan untuk memprediksi customer churn pada dataset Telco Customer Churn. Hasil perbandingan model menunjukkan

bahwa Random Forest memberikan performa terbaik dengan nilai accuracy sebesar 84,88%, precision sebesar 83,24%, recall sebesar 87,34%, F1-Score sebesar 85,24%, dan ROC-AUC sebesar 92,94%.

Hasil analisis karakteristik pelanggan menunjukkan bahwa pelanggan yang menggunakan kontrak bulanan, memiliki biaya layanan yang lebih tinggi, serta memiliki masa berlangganan yang relatif singkat cenderung memiliki risiko churn yang lebih besar dibandingkan kelompok pelanggan lainnya. Selain itu, berdasarkan hasil analisis feature importance, variabel Contract, MonthlyCharges, tenure, AvgChargePerMonth, dan TotalCharges teridentifikasi sebagai faktor-faktor yang memberikan pengaruh paling besar dalam proses prediksi customer churn.
Secara keseluruhan, model yang dikembangkan mampu menunjukkan kinerja klasifikasi yang baik sekaligus menghasilkan informasi yang bernilai untuk mendukung implementasi Customer Intelligence. Informasi tersebut dapat dimanfaatkan oleh perusahaan untuk mengidentifikasi pelanggan yang berpotensi melakukan churn, sehingga strategi retensi dapat dirancang dan diterapkan secara lebih tepat sasaran serta didasarkan pada hasil analisis data.

Saran

Penelitian pada masa mendatang dapat memanfaatkan dataset dengan ukuran yang lebih besar atau menggunakan data yang berasal dari berbagai sektor industri agar model yang dihasilkan memiliki kemampuan generalisasi yang lebih baik. Selain itu, penerapan algoritma lain, seperti XGBoost, LightGBM, atau CatBoost, dapat menjadi alternatif yang layak dipertimbangkan untuk meningkatkan performa prediksi customer churn.
Pengembangan lanjutan juga dapat diarahkan pada integrasi model prediksi ke dalam dashboard interaktif maupun sistem pendukung keputusan. Dengan pendekatan tersebut, hasil prediksi dapat dimanfaatkan secara langsung oleh perusahaan untuk melakukan pemantauan pelanggan secara lebih efektif serta mendukung penyusunan dan penerapan strategi retensi pelanggan yang berkelanjutan.

E. DAFTAR PUSTAKA

Ahmad, A. K., Jafar, A., & Aljouie, A. (2023). Customer Churn Prediction Using Machine Learning Approaches. *IEEE Access*.

Alotaibi, F., & Haq, E. U. (2024). Customer Churn Prediction for Telecommunication Industry Using Machine Learning and Ensemble Methods. *Engineering, Technology & Applied Science*

Breiman, L. (2023). *Random Forests*. Springer.

Fernandez, A., Garcia, S., & Galar, M. (2022). *Learning from Imbalanced Data Sets*. Springer.

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Huang, Y., Zhang, H., & Li, X. (2023). Customer Churn Prediction Using Machine Learning Techniques. *Electronics*.

Ullah, M. (2022). Customer Churn Prediction Using SMOTE and Machine Learning. *Applied Sciences*.